Day 2: The primary archival Databases

Bioinformatics ***Baxevanis and Oullette eds. Chapter 3 and 4. pp 45-81***

Primary and Derivative Databases

Primary sequence databases contain original reports of biological sequences from the investigators that determined them. Typically there are no additional data or interpretation added by the maintainers of these databases. These collections are archives of sequence information in that they contain records that may be unmodified or updated from the time they were submitted. A second consequence of this archival nature is redundancy in the data; there may be many examples of the same target sequence in the database submitted by different researchers.

The earliest examples of primary sequence databases are protein sequence collections. The Protein Information Resource (PIR) at Georgetown University is a direct descendant of one of the early protein sequence collections. As DNA cloning and sequencing technology improved, the number of nucleotide sequences available quickly surpassed directly determined amino acid sequences. Today by far the largest primary sequence databases are nucleic acid sequence databases. In fact the majority of protein sequences available today are based on conceptual translations of the coding regions on DNA sequences. This is true even in protein databases like PIR. Therefore we will treat protein sequence databases as derivative databases, to be covered on day 3.

The International DNA Sequence Database Collaboration

The most important primary DNA sequence databases today are the three members of the international DNA sequence collaboration: GenBank, the European Molecular Biology Laboratory (EMBL) database and the DNA Database of Japan (DDBJ). All three of these databases accept direct submissions of sequence data and are products of government-sponsored institutions in their respective countries. GenBank is produced and maintained by the National Center for Biotechnology Information at the National Institutes of Health in the United States, the EMBL database by the European Bioinformatics Institute in the United Kingdom and DDBJ by the Center for Information Biology of the National Institute of Genetics in Japan. All of these entities maintain a presence on the World Wide Web that includes browser-based access to data and tools for sequence analysis. The scope of these centers includes more than just their primary database product. All are active centers of computational biology and bioinformatics research and produce other data products including many important derivative databases.

The discussion here will focus on GenBank. Most of most of what is said will apply to the other two databases as well.

GenBank

The GenBank database has its origins in the dim past when it was produced in bound volumes. As the number of sequences increased and computer technology advanced, the database was made available on CD-ROM and came with software for accessing the data. The CD-ROM version was discontinued in 1997 when the number of CDs required became prohibitive. Right now GenBank is available through the Internet on the NCBI ftp site (URL: ftp://ftp.ncbi.nlm.nih.gov/genbank/). On the NCBI ftp server the database is made available as full data releases every two months in the even numbered months of the year. Between releases daily updates are provided. For each release, important information including release statistics is in the Release Notes (URL: ftp://ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt). The current release (123) contains over 12 billion bases and more than 11 million sequences from over 80,000 species.

Data files and GenBank Division Codes

On the ftp site the GenBank data are divided into series of sequence files. Originally each of these sequence files corresponded to one GenBank division. The GenBank divisions are identified by a three-letter division code, for example the BCT (bacterial) division or the PRI (primate) division. The days of one file per division are long past now; the EST division is split into more than 100 files. However all of the more than 11 million sequences in GenBank are still separated into a handful of divisions. A discussion of GenBank divisions is a helpful in describing the kinds of data in GenBank and is also useful in searching the data on the NCBI web site using the Entrez system. For the purposes of this discussion we'll recognize two kinds of GenBank divisions, traditional and special sequence divisions.

Traditional Divisions

The traditional GenBank divisions contain sequences that are determined to a high degree of accuracy (1 error in 10000) and often have extensive annotation about the biology or features of the sequence. At first glance these traditional divisions appear taxonomic in nature. Closer inspection shows the overriding purpose in establishing them initially was to create single files of reasonable size and taxa were spilt or lumped to accomplish this. For example, the primate (PRI) and rodent (ROD) sequences were separated from the rest of the mammalian sequences (MAM) because there were a large number of primate and rodent sequences. On the other hand the fungal and plant sequences were lumped into the PLN division because originally there were fewer of these. Sequence data can be submitted to these divisions through a web based form called BankIt (URL: http://www.ncbi.nlm.nih.gov/BankIt/) or for more complex submissions can be prepared using the NCBI standalone tool Sequin (URL: http://www.ncbi.nlm.nih.gov/Sequin/index.html)

Special Divisions

With changes in DNA sequencing technology and strategies, a number of special GenBank divisions were established. These are not based on the source organism of the sequence but are based on the technique used to generate the sequence or the intended use of the sequence. A unifying characteristic of these divisions is that they tend to be submitted in large batches by single projects. The GSS, EST and STS sequences also exist NCBI databases apart from GenBank: dbGSS, dbEST and dbSTS. The format of the records within these databases is quite different than that used in GenBank. Submissions to these divisions are handled through different procedures and different staff than traditional submissions.

First pass sequence divisions

The expressed sequence tag (EST) division and the genome survey sequence division (GSS) were established to hold first pass single read sequences that have little or no annotation. Because these data are single sequence reads, the amount of sequence in each record is limited, and is likely contain sequencing errors including frame shifts and base miscalls.

EST division

The EST division holds automatically generated partial cDNA sequences. These sequences are derived from arrayed cDNA libraries. For each clone in the library, only a single read is obtained from each end of the insert using the standard sequencing primers. Thus there can be two sequences in GenBank for each clone. In the case of directionally cloned libraries these will be the 5' end of the cDNA and the 3' end. Robots are often used to automate the process of sequencing these clones. Large numbers of clones can be partially sequenced very rapidly using this strategy. Good sets of EST data are available for a number of organisms. In fact the EST division is the largest division of GenBank. Although largely unannotated and error prone, these data provide a rich source of information about the expressed sequences in a particular cell type tissue or ultimately the organism. The EST data are an important resource gene discovery and gene expression data. At the NCBI, the EST data have been used to generate a derivative database, UniGene that attempts to organize these data into gene based clusters. We will discuss UniGene in more detail on day 3 of the course.

GSS division

The GSS division contains data that are the genomic equivalent of the EST data. That is first pass single reads of genomic clones. The bulk of the data in the GSS division are derived from bacterial artificial chromosome (BAC) libraries. BACs are the large insert genomic clones that are used in complex genome projects like the human genome project. This is explained in more detail below when we describe the HTG division. Sequencing centers will produce preliminary reads for these clones sometimes as a prelude to producing more complete sequences. These survey go in the GSS division. Another related category of sequence comes form the extension of sequencing primers onto the insert of the clone. These so called BAC end sequences are used to identify overlapping clones and creating tiling paths for assembling large genomic contigs. The GSS division also contains whole genome shotgun sequencing reads for some organisms notably certain protozoan parasites. These GSS sequences are important resources for genomic sequence for these organisms even in this unassembled form.

The High Throughput Genome Sequence Division

Many of the large-scale genome-sequencing projects such as the publicly funded human genome project rely on a strategy that has been called hierarchical shotgun sequencing. Genomic libraries are made in large insert BAC vectors, which have an insert capacity of around 150 Kb. The clones from these libraries are arrayed and then subcloned into plasmid vectors. The resulting mini libraries are then randomly sequenced until enough sequence is obtained at high accuracy to assemble this shotgun sequence to generate the insert sequence of the clone. Even the early stages of this assembly process are useful. So that investigators can have access to these incomplete or draft sequences GenBank established the High Throughput Genome sequence (HTG) division. Within the draft or HTG sequences, GenBank recognizes different phases of completion. These phases are based on the degree of coverage and assembly of the sequence in the record. Phase 1 records have sufficient coverage to have several assembled regions (contigs). However the order and orientation of these is unknown and there will still be gaps of unknown length in the sequence. Phase 2 records have progressed to the point that the order and orientation of the assemblies is known, but there are still gaps. As more sequence becomes available the submitters update the records and the records will progress through the draft phases until the coverage and accuracy are sufficient for the sequence to move to phase 3. At that point the record moves from the HTG division into one of the traditional GenBank divisions: A human sequence would move to the PRI division. A fly sequence would move to the INV division. A zebrafish sequence would move to the VRT division. Even though the sequences within them are incomplete, the draft sequences are still useful. The NCBI assembly of the human genome depends on draft sequences; about half of the human genome is still in the HTG division.

The Sequence Tagged Site (STS) Division

Records in the STS division are essentially mapping reagents. A sequence tagged site is essentially a recipe for amplifying a specific fragment of genomic DNA using the polymerase chain reaction (PCR). The records generally include a pair of primers and the sequence of genomic DNA they amplify. STS markers are designed based on the sequence of a known gene, an EST, an mRNA or genetic marker. STSs are used in the technique of radiation hybrid (RH) mapping as a means of constructing a physical map of a genomic region. In RH mapping a cell line from the species of interest (human for example) is given a lethal does of radiation. One effect of this is to break the genome into fragments. The fragmented genomic DNA of the irradiated cells can be rescued by fusing the cells with those of a different species. The resulting hybrid cells variously retain and expel fragments of the foreign genome so that unique clones from the hybrid line have differing portions of the foreign genome. Genomic DNA isolated from these clones can then be tested by PCR with STS markers to the irradiated genome. The probability that marker occur together in the same hybrid clone is inversely related to the distance between them in the original genome. This is in essence like a genetic recombination experiment. The difference is that the breakpoints between markers are caused by radiation. Since the genetic position of many of these markers often known, radiation hybrid map positions can be integrated with genetic maps. Finally since STS markers are also sequenced based markers they can be mapped onto the assembled genomic sequence. The NCBI tool electronic PCR (ePCR) will search a sequence for the presence of markers from the STS division. This tool has been important in assembling the human genome sequence.

Other Special Divisions

The patent division (PAT) contains sequences provided by the US Patent and Trademark office. These sequences are not well annotated and not particularly useful even for patent claim investigation since GenBank cannot assure that this divison includes all patents.

The contig division (CON) contains records that are instuction sets for assembling larger sequences. This division exists partly because GenBank has a 350 Kb limit for a single sequence. An example of a record in the CON division is the one containing instructions for assembling the *Escherichia coli* K12 genome from the < 350 Kb pieces in the BCT division.

The high throughput cDNA (HTC) division was recently created for draft cDNA records. Like the HTG division these sequences can be finished and then will move into the corresponding traditional division.

The GenBank Flatfile Format

If we were to download one of the sequence files from the genbank directory on the NCBI ftp site and look inside, we would find that it is a large text file. Within that file are many individual records, one following right after another. The only thing separating the records is a special mark consisting of a double front slash "//" at the bottom of the sequence. This kind of a large text file database containing multiple records is known generically as a flatfile database. Hence The format of the records within that GenBank sequence file has come to be known as the GenBank flatfile format. This format is discussed in detail in **Bioinformatics** Baxevanis and Oullette eds. Chapter 3. pp.49-58. We will follow that discussion closely so I won't reproduce it here. We will be using the record included on the next page as an example.

```
LOCUS       AF062069     3808 bp    mRNA            INV       02-MAR-2000
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
KEYWORDS    .
SOURCE      Atlantic horseshoe crab.
  ORGANISM  Limulus polyphemus
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
            Xiphosura; Limulidae; Limulus.
REFERENCE   1  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
  JOURNAL   J. Neurosci. (1998) In press
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
FEATURES             Location/Qualifiers
     source          1..3808
                     /organism="Limulus polyphemus"
                     /db_xref="taxon:6850"
                     /tissue_type="lateral eye"
     CDS             258..3302
                     /note="N-terminal protein kinase domain; C-terminal myosin
                     heavy chain head; substrate for PKA"
                     /codon_start=1
                     /product="myosin III"
                     /protein_id="AAC16332.2"
                     /db_xref="GI:7144485"
                     /translation="MEYKCISEHLPFETLPDPGDRFEVQELVGTGTYATVYSAIDKQA
                     NKKVALKIIGHIAENLLDIETEYRIYKAVNGIQFFPEFRGAFFKRGERESDNEVWLGI
                     <sequence omitted>
                     LIRQFGFARRISFVDFLNRYQYLAFDFNENVELTKENCRLLLLRLKMDGWTLGKNKVF
                     LKYYSEEYLSRIYETHIKKIVKVQAIARKYFVKVRQSKTKPH"
BASE COUNT     1201 a    689 c    782 g   1136 t
ORIGIN
        1 tcgacatctg tggtcgcttt ttttagtaat aaaaaattgt attatgacgt cctatctgtt
       61 gttgtgttac acaggtacat attaataaca ggtagctaac gtacttatat atacatatat
          <sequence omitted>
     3781 aagatacagt aactagggaa aaaaaaaa
//
```

Exercises

1. Go to the GenBank directoy on the NCBI ftp site. Find the current release statistics. Enter the directory containing all prevous release notes. Find the release number and date when GenBank was about half its current size.

2. Use the search GenBank box on the NCBI web page to retrieve the CON division record for the *Mycoplasma pneomoinae* complete genome. (Enter the accession number, U00089, and hit go). How many GenBank records make up this genome?

3. Use the same procedure used in 3 above to retrieve the human HTG record AC013402. How many unordered pieces are in the record now. How many times has this record been updated since it first appeared? Trace its history all the way back to the first version. Based on the update date about when did this record first appear? How many unordered pieces were there then? Now use electronic PCR, linked to the hotpots on the NCBI home page, to identify the STS markers present in this record. How many are there? These include RH markers and genetic markers. Which one is also a genetic marker?

4. The are many different kinds of sequence formats (e.g. GenBank, FASTA, GCG). Some sequence analysis packages are quite picky about the formats they accept. The utility readseq converts between many common formats. This can be run from the command line by typing `readseq.` Use readseq to convert GenBank format of AF062069 to FASTA, GCG and EMBL formats.

5. Use the search GenBank box to retrieve the record containing the *Entamoba dispar* record AF118397. Display the record in FASTA format and save it to your home directory. Use nedit or pico to edit the sequence and remove the information from the title of the record. Launch Sequin and prepare a submission using this sequence. Annotate a gene, protein and coding region feature on the record. Use the built in Open Reading Frame Finder (ORF Finder) to identify the location of the coding region.